

901 809

NAWCWPNS TP 8072

AD-A271 660



Automatic Target Recognition Display Format Study

by
Edward D. McDowell
Oregon State University
for
Aircraft Weapons Integration Department (Fighter/Attack)

NOVEMBER 1992

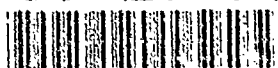
NAVAL AIR WARFARE CENTER WEAPONS DIVISION
CHINA LAKE, CA 93555-6001



Approved for public release; distribution is unlimited.

DTIC
ELECTE
OCT 28 1993
S D

93-25970



93 10 26 004

Naval Air Warfare Center Weapons Division

FOREWORD

This report covers work performed in fiscal years 1990 to 1992. This work was supported by Navy Exploratory Development Program, Command Systems (N02C) Technology Block, Human Factors Task 7 and by the U.S. Navy - American Society for Engineering Education (ASEE) Summer Faculty Research Program.

This work was reviewed for technical accuracy by Marion P. Kibbe and Jan S. Stiff.

Approved by
M. K. BURFORD, *Head*
Aircraft Weapons Integration Department (Fighter/Attack)
26 October 1992

Under authority of
W. E. NEWMAN
RAdm., U.S. Navy
Commander

Released for publication by
W. B. PORTER
Deputy Commander for Research & Development

NAWCWPNS Technical Publication 8072

Published by..... Technical Information Department
Collation..... Cover, 17 leaves
First printing... 90 copies

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE November 1992	3. REPORT TYPE AND DATES COVERED Summary, 1991-92
4. TITLE AND SUBTITLE Automatic Target Recognition Display Format Study			5. FUNDING NUMBERS
6. AUTHOR(S) Edward D. McDowell			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Air Warfare Center Weapons Division China Lake, CA 93555-6001			8. PERFORMING ORGANIZATION REPORT NUMBER NAWCWPNS TP 8072
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)			10. SPONSORING/MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES			
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.			12b. DISTRIBUTION CODE
13. ABSTRACT (Maximum 200 words) See reverse.			
14. SUBJECT TERMS Aided target recognition Automatic target recognition Man-machine interface ATR ATR Decision support systems (targeting) Synergism ATR/MMI Interface with ATR Target cueing Targeting			15. NUMBER OF PAGES 33
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT SAR

Block 13. Contd.

In the foreseeable future, automatic target recognition (ATR) systems are going to become an integral part of targeting systems. How these ATRs interact with operators is clearly going to influence overall system accuracy. The objective of the two studies covered by this report was to explore the effect of two specific ATR-operator interface factors on identification accuracy. The two factors were the number of recommendations the ATR provides the operator and the nature of the accuracy measure figure of merit, (FOM) provided with each recommendation. A simulated range only radar (ROR) task was used to explore the effect of these two factors. Two levels (1 and 5) were used for the number of recommendations and three levels (qualitative, quantitative, and none) were used for the accuracy measure. The effect of these factors was investigated under two levels of ATR and operator accuracy, nominally 80 and 40%, respectively. These factors influenced both the likelihood that an operator would alter a correct ATR recommendation and the likelihood that an operator would correct an incorrect ATR recommendation. The influence was greater in the former than in the latter. Increasing the number of recommendations significantly increased the likelihood that an operator would disagree with the ATR's first recommendation, regardless of whether this first recommendation was correct. The type of FOM used had only a small effect when associated with a single recommendation. However when paired with five recommendations, subjects were less likely to disagree with a recommended identification if a quantitative FOM was provided.

CONTENTS

Introduction	3
Background	4
Methodology	5
General	5
Images	5
Information Quality	6
Display Formats	6
Subjects	8
Procedure	8
Results	9
Overall Accuracy	9
Independence	10
Ships	11
Aiding	12
Overall Mean Response Times	13
Display Format	14
Summary and Discussion	22
Recommendation	23
Appendix: Agreement-Disagreement and Signal Detection Theory	25
References	32

DTIC QUALITY INSPECTED 3.

Accession For	
BTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

INTRODUCTION

As the range of air-to-surface weapon systems steadily increases, target identification, as required by the rules of engagement, is becoming more difficult. Positive target identification can no longer be based solely upon direct visual contact. Rather, target identification will increasingly be based upon information acquired by a suite of passive and active sensors.

As envisioned, future targets will be detected by a set of sensors. The output from this sensor suite would be used by both an automatic target recognition (ATR) system and an operator. Using appropriate pattern recognition algorithms, the ATR would determine the most probable target identity, which would also be displayed to the operator. This general model is illustrated in Figure 1.

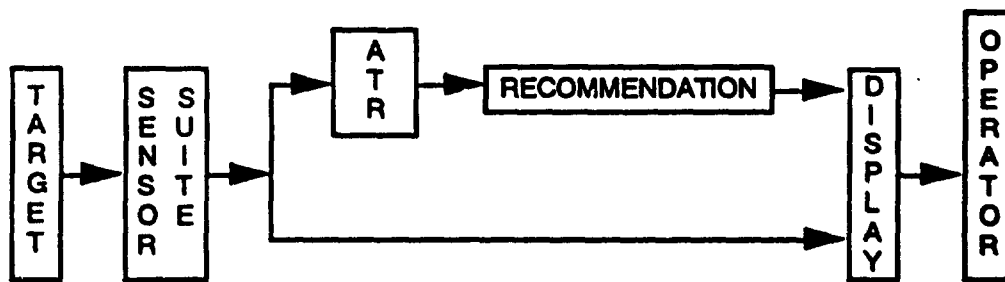


FIGURE 1. Model of an Operator/ATR System.

In this paradigm, the operator has both the ATR's recommendation and the original sensor data available for target identification. Ideally, the operator would combine this information in an optimal manner.

The operator could follow various procedures in making the final identification decision. As an example, the operator might (1) consult the sensor data and formulate a hypothesis regarding the target's identification, and (2) consult the ATR's recommendation. If recommendations (1) and (2) concur, then that recommendation would become the final identification. If the two do not concur, then the operator must resolve the conflict by selecting the identification hypothesized by himself, the one selected by the ATR, or a third alternative.

The level of performance actually achieved in an operator/ATR system is a function of several factors, not the least of which is the nature of the interface between the ATR and the operator. Clearly the nature of the information an ATR provides could significantly influence the accuracy of the overall operator/ATR system.

The purpose of the set of two studies covered in this report was to investigate several questions regarding the display of ATR information. One objective was to determine whether system performance would vary with the number of ATR provided recommendations. A second objective was to evaluate the effectiveness of various types of figures of merit (FOMs) used to characterize the accuracy of each recommendation.

BACKGROUND

Several studies have examined the performance of operators in single or multiple sensor systems (References 1 through 4). Generally these studies have concentrated on an operator's ability to utilize all available information. In a targeting performance study using multiple imaging sensors, Kibbe and Weisgerber observed that "... where the quality of information from one sensor was worse than that provided by the other sensor, accuracy may be worse than it would have been if only the better of the two sensors had been used" (Reference 2). Consequently, they concluded that operators do not always make the most effective use of available information in a multisensor environment and that it may not always be advisable to make all the information available to the operator.

In another study in which they investigated the interaction between an ATR and simulated forward-looking infrared (FLIR) imagery, Kibbe and Weisgerber found that an operator was effective in integrating image and ATR information (Reference 5). They concluded that an operator should be an integral part of the system because an operator/ATR system performed better than either the operator alone or the ATR alone. In contrast, Foyle, in a sensor fusion study using simulated ROR and FLIR imagery, found that an operator using both sensors sometimes performed better and sometimes worse than either sensor alone (Reference 3).

Weisgerber and Savage found that operators could derive useful information from even a low reliability ATR (Reference 6). They also concluded that in most instances the accuracy of the operator/ATR system was superior to either the operator or the ATR alone. However, Adams found that operator/ATR system accuracy was inferior to that of a highly accurate ATR alone and about equal to an inaccurate ATR (Reference 4).

Although providing valuable insight into the relationship between an operator and an ATR, these studies appear to arrive at conflicting results, particularly with regard to system performance as a function of ATR and operator performance.

METHODOLOGY

GENERAL

This report describes the results of two studies, one conducted at the Naval Air Warfare Center Weapons Division (NAWCWPNS), China Lake, Calif., referred to as study 1, and the second conducted at Oregon State University, and referred to as study 2. The same general methodology was utilized for both studies.

The general approach for these two studies was to simulate a ship identification task and using this task, to test the effectiveness of various formats for displaying ATR information. The simulation was implemented on an IBM personal computer (PC) compatible machine.

IMAGES

Fifteen ships were selected from *Jane's Fighting Ships* (Reference 7) to serve as the set of possible target ships. Each of the 15 ship silhouettes was represented by 60 vertical bars. Examples are shown in Figure 2. Names were then associated with each image. No attempt was made to associate a ship's actual name to its image. Images were distorted by adding normally distributed deviations to each vertical bar. Each of the 60 deviations used in an image were randomly selected from a normal population with a zero mean. The deviations' standard deviations were adjusted to give the operator and ATR reliability levels desired. If a resulting bar height was negative, then that bar was displayed as a zero height bar.

INFORMATION QUALITY

Two levels of image distortion were selected for the experiments. These were selected to yield autonomous operator and ATR accuracies of 80 and 40%, respectively. The 80% accuracy is referred to as distortion level 1 (DL1), the low distortion level; and the 40% accuracy level is referred to as distortion level 2 (DL2), the high distortion level. The quality of the information provided to the ATR and the operator was co-varied.

DISPLAY FORMATS

The two display format characteristics varied in these experiments were the number of recommendations displayed (ND) and the type of FOM (the reliability measure) displayed with each recommendation. The two levels selected for the number of recommendations were one and five.

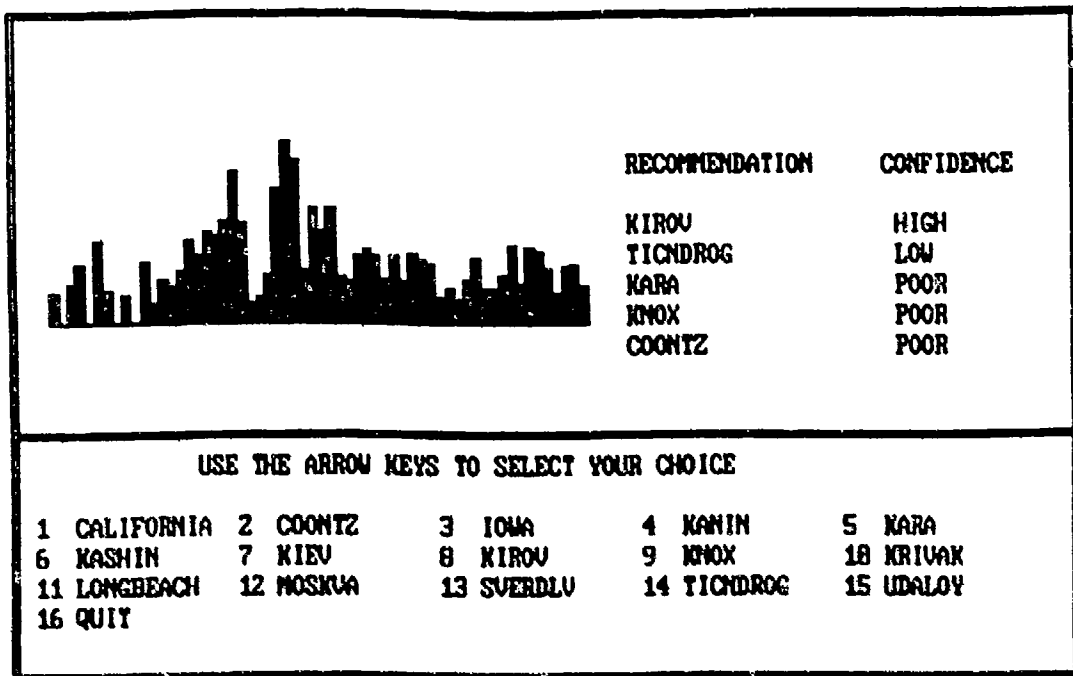
Because it is difficult or impossible to determine an accurate ATR probability, ATR designers are particularly interested in display formats that either do not include a FOM or use a qualitative or a non-probabilistic quantitative FOM.

Three levels were used for the FOM type: qualitative (q), quantitative (Q), and none (N). For the qualitative level, one of the four words (high, medium, low, or poor) appeared with each recommendation. "High" indicated the recommendation had a 0.9 or higher probability of being correct, "medium" a 0.8 to 0.9 probability, "low" a 0.7 to 0.8 probability, and "poor" less than 0.7 probability. An example of a qualitative FOM display is shown in Figure 2a.

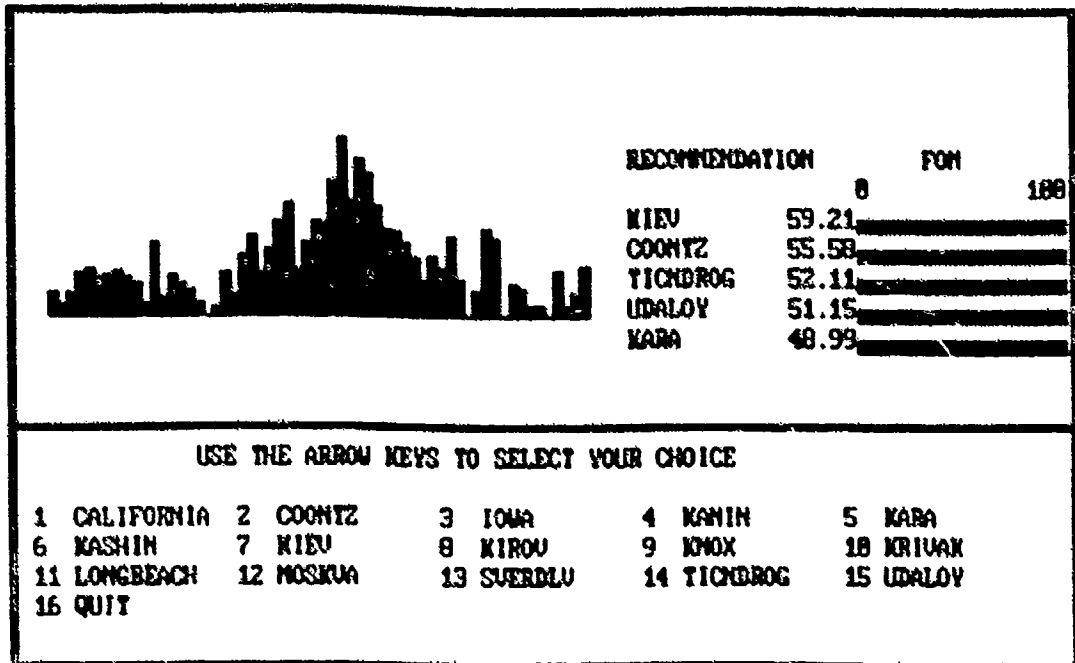
The quantitative FOM used was a numerical value scaled to range from 0 to 100. A horizontal bar, also scaled from 0 to 100, was displayed with each recommendation. Higher values indicated more reliable recommendations. The quantitative FOMs did not represent probabilities. For example, a FOM of 62 corresponded to a reliability of approximately 0.9, a FOM of 56 a reliability of 0.8, and an FOM of 53 a reliability of 0.7. This display format is illustrated in Figure 2b.

A recommendation without a FOM was the third display possibility. For the reasons cited above this alternative was of particular interest to the ATR designers.

Crossing the three levels of FOM with the two levels of ND resulted in six display formats. A seventh alternative, no recommendation (C0), was added as a control. The seven levels of display format are named and summarized in Table 1.



(a) Five recommendation qualitative FOM display.



(b) Five recommendation quantitative FOM display.

FIGURE 2. Example Display Format.

TABLE 1. Display Formats.

Treatment	Display type	Number of recommendations	Type of FOM
C0		None	N/A
1q	1	One	Qualitative
1Q	2	One	Quantitative
1N	3	One	None
5q	4	Five	Qualitative
5Q	5	Five	Quantitative
5N	6	Five	None

SUBJECTS

Fourteen subjects completed each study. For study 1, the fourteen subjects were volunteer civilian employees of NAWCWPNS. For study 2, the fourteen subjects were paid Oregon State University undergraduate students.

PROCEDURE

The general nature of the experiment was described to each subject, who was then asked to sign a consent to serve form. Each subject was then instructed in how to start the test, and was provided a hard copy of the 15 undistorted ship silhouettes to use for reference throughout the experiment.

Detailed instructions were integrated into the computer program. The instructions described the ATR display to be used, and whether it was a practice block (which included feedback) or test block (which did not). The instructions also reminded the subject of the reliability of the ATR. That is, if the quantitative ATR was to be used, the instructions reminded the subject that recommendations with FOMs of 62 would be correct on about 90% of the trials.

For each trial a cathode ray tube (CRT) displayed an image and, when appropriate, an ATR recommendation. The 15 possible ship names were displayed across the bottom of the screen. The keyboard arrow keys were used to move the cursor to the correct ship name and then the "enter" key was pressed. Once the "enter" key was pressed, the subject's response was recorded, and the program automatically proceeded to the next trial. No time limits were imposed for study 1. However, for study 2, a 30-second time limit was enforced: after 20 seconds, a warning message indicating the

time remaining was displayed on the screen. This message counted the number of remaining seconds down to 0.

In order to become acquainted with the identification task, each subject began with a training session. During the training session, the subject went through the 15 undistorted silhouettes twice. The training session was followed by an initial practice block where the subject completed 30 trials without the aid of an ATR but with feedback.

After the training and initial practice, pairs of practice and test blocks were presented. Both practice and test blocks consisted of 30 trials. The subject always received a practice block with a particular ATR format just prior to its corresponding test block. The order of ATR formats was counterbalanced. Thus, there were seven sequences of blocks. During each block, each ship was presented as the target at each of the two distortion levels. The order of distortion levels and target ships was randomized for each subject and block. In an attempt to mitigate the effects of fatigue, the testing was distributed over two test sessions, with never more than one intervening day between sessions. The first session consisted of the training and practice without an ATR (C0), and three of the seven ATR display combinations. The remaining four display formats were presented during the second session. Two subjects were tested for each counterbalancing sequence.

For each trial, the subject's response, the correct response, the ATR's first recommendation, and the subject's response time were recorded for analysis.

RESULTS

OVERALL ACCURACY

The image distortion levels had initially been selected to give autonomous accuracies of 80 and 40% for the ATR and operator, respectively. An examination of the data revealed that the ATR's accuracy was consistent across the two studies. At the low distortion level, the ATR was correct on 84.6% for study 1 and 85.3% for study 2, resulting in an overall accuracy of 84.9%. For the high distortion level, the ATR was correct on 39.9 and 41.4% of the trials, respectively, for an overall accuracy of 40.7%.

In the two studies, the subjects' unaided accuracy averaged 56.1 and 51.8% for the low distortion level and 39.9 and 39.4% for the the high distortion level, respectively.

Overall, the subjects' unaided accuracies closely approximated the 40% targeted for the high distortion level, but were significantly under the 80% targeted for the low distortion level. At the low distortion level, the subjects' unaided accuracies were 28.4 and 33.4% less than the ATR's. At the high distortion level, the subjects' unaided accuracy was 0.0% higher and 1.9% lower than the ATR's, respectively, for the two studies.

INDEPENDENCE

Because both the subjects and the ATR were provided with the same imagery, it could be conjectured that their performance would be highly correlated. That is, it was hypothesized that the subjects would have difficulty classifying the same images that the ATR would have difficulty classifying.

Although the subjects did not have access to its recommendations, the ATR was fully operational and its recommendations were being recorded during control test block trials (C0). In order to test the null hypothesis that the subjects and the ATRs were operating independently, correct and incorrect responses during this test condition (C0) were cross-tabulated for each distortion level and study. The results of this tabulation are shown in Table 2, Columns A, B, C, and D.

TABLE 2. Analysis of Condition (C0).

		FREQUENCY FOR STUDY 1				FREQUENCY FOR STUDY 2			
		A. LOW DIST.		B. HIGH DIST.		C. LOW DIST.		D. HIGH DIST.	
		ATR ID		ATR ID		ATR ID		ATR ID	
SUBJ	ID	W	C	W	C	W	C	W	C
		14	78	78	48	18	83	83	44
		C	14	104	50	34	13	96	45
CHI-SQ =		0.50		0.12		1.45		2.62	
SIG. LEVELS =		0.473		0.73		0.23		0.11	
W = INCORRECT RESPONSE and C = CORRECT RESPONSE									
Column A of Table 2 should read: Seventy-eight ship images were identified correctly by the autonomous ATR but incorrectly by the unaided subjects. Fourteen ships were identified incorrectly by both the ATR and the subjects acting alone, 14 were identified incorrectly by the autonomous ATR and correctly by the unaided subjects, and finally 104 ships were identified correctly by the autonomous ATR and the unaided operators.									

Chi-square contingency tests were used to compare these observed frequencies to the frequencies that would be expected if the ATR and the subjects were performing independently. No chi-square values were significant, indicating that one cannot reject the null hypothesis that the subjects and the ATR were operating independently.

Given that the two were provided the same information, this is an interesting result. It would suggest that even though both the subjects and the ATR are using the same information, they are using it in quite different classification algorithms.

SHIPS

The proportions of responses with which each ship was correctly identified by the subjects and the ATR during the unaided control trials (C0), at each of the distortion levels, are shown in Table 3. The results in Table 3 were obtained by averaging across both studies. The hypothesis that each of the 15 ships could be identified by the subjects with equal accuracy was tested using a chi-square contingency table. This hypothesis was rejected for both distortion levels and studies ($\alpha \leq 0.05$). Similarly, the hypothesis that the ATR could identify the ships with equal ease was tested. This hypothesis was also rejected for both distortion levels and studies ($\alpha \leq 0.05$).

The hypothesis that the subjects' distributions of accuracies across ships were the same across the two studies and across the two distortion levels was tested. In both cases, there was insufficient evidence ($\alpha > 0.05$) to reject the null hypothesis of no difference. Specifically, the ships the subjects had difficulty identifying at the high distortion level were the same ships they had difficulty identifying at the low distortion level, and the ships that were difficult to identify in study 1 were also difficult to identify in study 2.

The ATR's accuracy was similar across both studies, but not across distortion levels ($\alpha < 0.05$).

TABLE 3. Subject and ATR Accuracy.

	SHIPS														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LOW DISTORTION															
SUBJ.	0.57	0.46	0.32	0.25	0.35	0.50	0.71	0.5	0.46	0.50	0.75	0.82	0.61	0.89	0.39
ATR	0.96	1.0	0.5	1.0	1.0	1.0	0.89	0.96	1.0	0.28	0.78	0.64	0.89	1.0	0.96
HIGH DISTORTION															
SUBJ.	0.42	0.21	0.18	0.18	0.25	0.5	0.36	0.25	0.32	0.42	0.71	0.67	0.5	0.61	0.36
ATR	0.18	0.89	0.0	0.57	0.71	0.64	0.14	0.25	0.78	0.0	0.0	0.0	0.50	1.0	0.17

An examination of Table 3 shows that, operating unaided and using the low distortion level imagery, subjects could identify only 2 of the 15 ships more reliably than the ATR. However at the high distortion level, this relationship was altered, with the subject being more reliable on 7 of the 15 ships. Although both the subjects and the ATR were nearly equal in average accuracy, (39.7 versus 39.0%), the ATR's accuracy was considerably more variable, failing to identify four ships correctly on even a single trial. The standard deviation of the subjects' accuracy across ships was 0.169, versus 0.344 for the ATR.

At the high distortion level, the ATR correctly reported ship no. 14 100% of the time that this ship was presented. This gives the ATR a perfect record in identifying ship no. 14, which appears impressive. However, the ATR also recommended ship no. 14 on 201 out of a possible 420 (47.85%) of the trials. Thus when reporting ship no. 14, the ATR was only correct 13.9% of the time, a less enviable record. This shows that the accuracy figures reported in Table 3 can be very misleading, and that errors as well as accurate identifications must be examined.

If the more reliable source, the operator or the ATR, is used for each ship, then at the low distortion level the average accuracy would be 88.4%, compared to 84.9% for the ATR alone. At the high distortion level, the comparable figures would be 55.1%, compared to 40.7%.

AIDING

As shown in Figure 3, when averaged across display formats, aiding positively affected overall accuracy. At the low distortion level, aided accuracy averaged 84.1 and 82.0%, respectively, for the two studies. Thus aided accuracy averaged 27.9 and 30.1% higher than the subjects alone, but 0.4 and 3.3%, respectively, less than the ATR's average accuracy.

At the high distortion level, aided accuracy increased to 55.6 and 51.0% correct, respectively, for the two studies. Thus aided accuracy exceeded ATR accuracy by 15.7 and 9.6%, respectively, and subject accuracy by a similar amount. Recall that when using the more reliable source, the estimated accuracy was 55.13%. Thus the subjects in study 1, when using an ATR, did slightly better than the score they would have obtained by simply picking the more reliable source for each ship, while the subjects in study 2 were somewhat below this level.

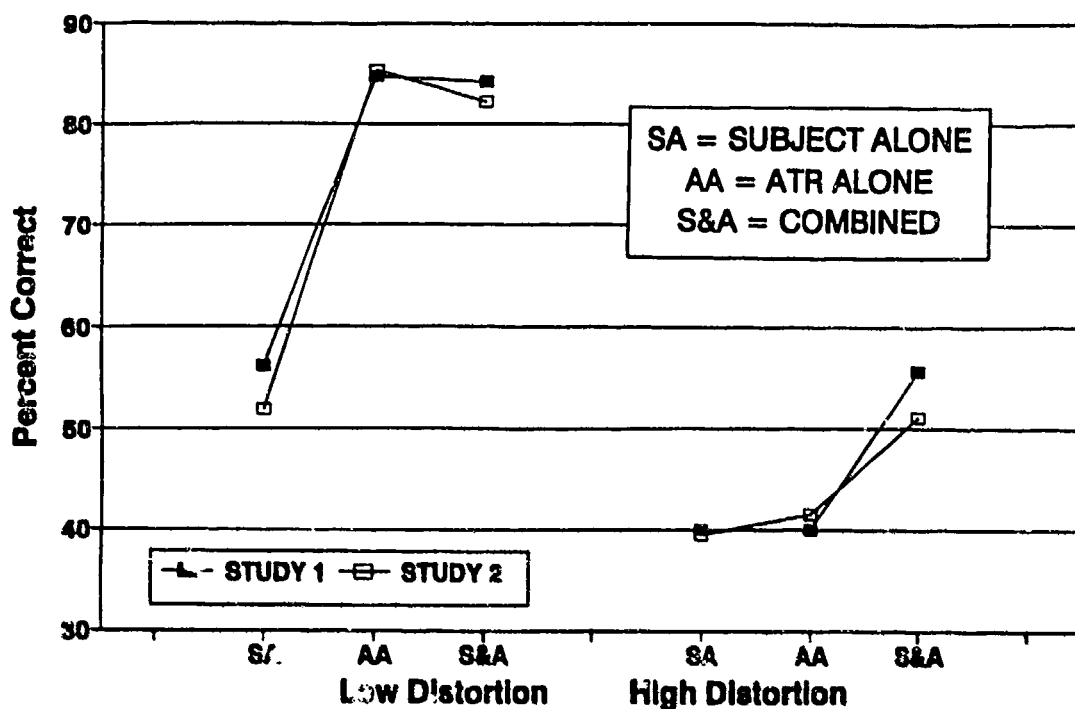


FIGURE 3. Aided Versus Unaided Performance Accuracy.

OVERALL MEAN RESPONSE TIMES

ATR aiding significantly reduced response times at the low distortion level. Without aiding, average subject response times at the low distortion level were 17.9 and 12.0 seconds for studies 1 and 2, respectively. With aiding, average response times decreased 4.4 and 3.0 seconds, resulting in mean response times of 13.4 and 9.0 seconds, respectively, for the two studies. An analysis of variance showed that both reductions were statistically significant ($\alpha \leq 0.01$).

However, as may be seen in Figure 1, aiding did not significantly affect mean response times at the high distortion level. For study 1, aiding increased mean response times from 19.1 to 20.7 seconds, while decreasing mean response times from 14.6 to 13.9 seconds for study 2. Neither change was statistically significant.

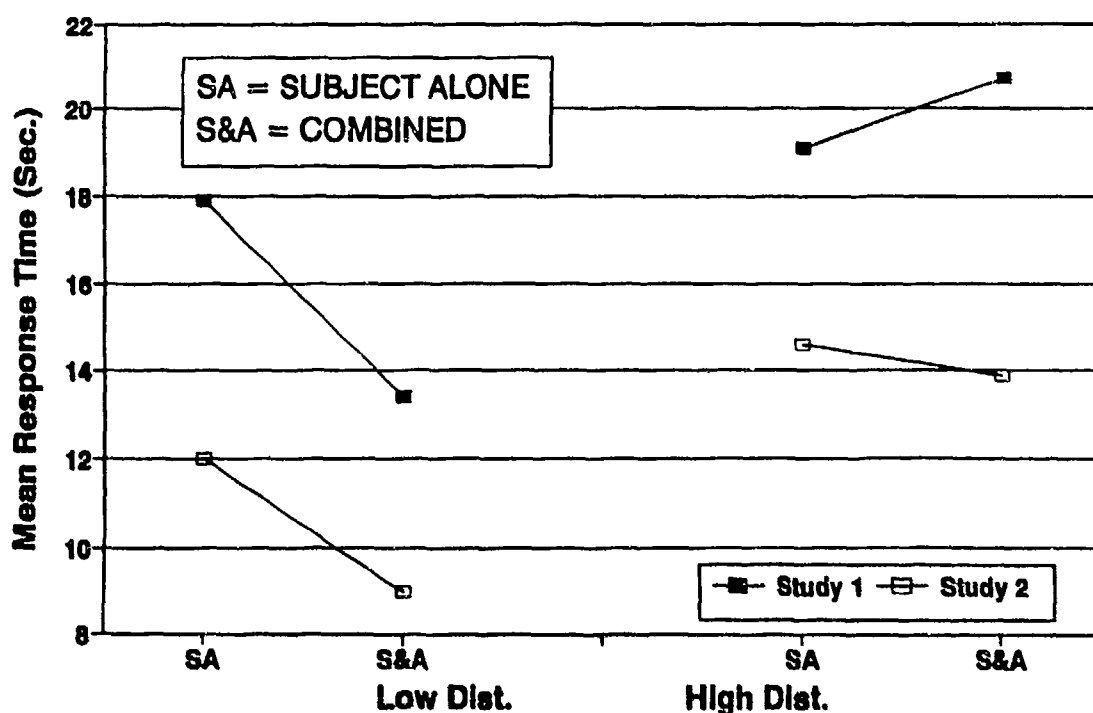


FIGURE 4. Mean Response Time for Aided Versus Unaided Performance.

Increasing image distortion increased mean response times for both studies, but more markedly for the aided condition. For the C0 conditions (shown in Figure 4 as SA), increasing image distortion increased mean response times from 17.9 to 19.1 seconds and from 12.0 to 14.6 seconds for the two studies, respectively. An analysis of variance showed that the increase was statistically significant for study 2 ($\alpha \leq 0.005$), but not for study 1. When an ATR was available, the increases in the mean response times as image distortion was increased were from 13.4 to 20.7 seconds and from 9.0 to 13.9 seconds, respectively. Both were statistically significant ($\alpha \leq 0.00001$).

DISPLAY FORMAT

General Model

In order to create a balanced model, the results for the control conditions (C0) were dropped from all further analyses. The statistical model used throughout the subsequent analysis treated subjects and test sequence as blocking factors and distortion level (low, high), the number of recommendations displayed (1, 5), and the FOM format (qualitative, quantitative, or none) as crossed factors.

Subjects and Order

Generally, both the subject and trial order factors were significant for all dependent measures for both studies, but because neither was of particular interest, a detailed discussion will be omitted.

Accuracy

The factors that significantly influenced system accuracy in both studies were for distortion level ($\alpha < 0.0001$ and $\alpha < 0.0001$), the distortion level by FOM interaction ($\alpha = 0.0315$ and $\alpha = 0.0647$), and the FOM by number of recommendations displayed interaction ($\alpha = 0.0449$ and $\alpha = 0.0126$). The FOM main effect ($\alpha = 0.0022$) and the distortion level by number of recommendations displayed interaction ($\alpha = 0.0040$) were significant in study 2 but not study 1.

The FOM by ND interaction is shown in Figure 5. When averaged across distortion levels and studies, the five recommendation-quantitative FOM display (5Q) was the most accurate. It averaged 73.1 and 71.2% accurate for the two studies, respectively, for an overall average of 72.1%. Conversely, the five recommendation-no FOM display (5N) was the least accurate for both studies, with average accuracies of only 66.2 and 60.2%, for an overall accuracy of 63.2%. Consequently, display formats with five recommendations were both the highest and lowest performing, resulting in a significant interaction between ND and FOM factors.

Format 5Q was clearly superior for the low distortion condition, while 5N was just as clearly inferior. The display formats were more equally matched when the images were highly distorted. This result is particularly interesting because the operator/ATR system performed better than either the operator or the ATR alone at this distortion level.

Thus, the observed overall accuracy differential can be attributed primarily to differences at the low distortion level, the level where the operator/ATR system performed no better than the ATR alone. At the high distortion level, where the system performed better than either the ATR alone or the operator alone, the differences due to display formats were smaller. Whether this observed difference in accuracy may be attributed to the ATR's high accuracy, the differential in performance between the subjects and the ATR, or both, is unclear.

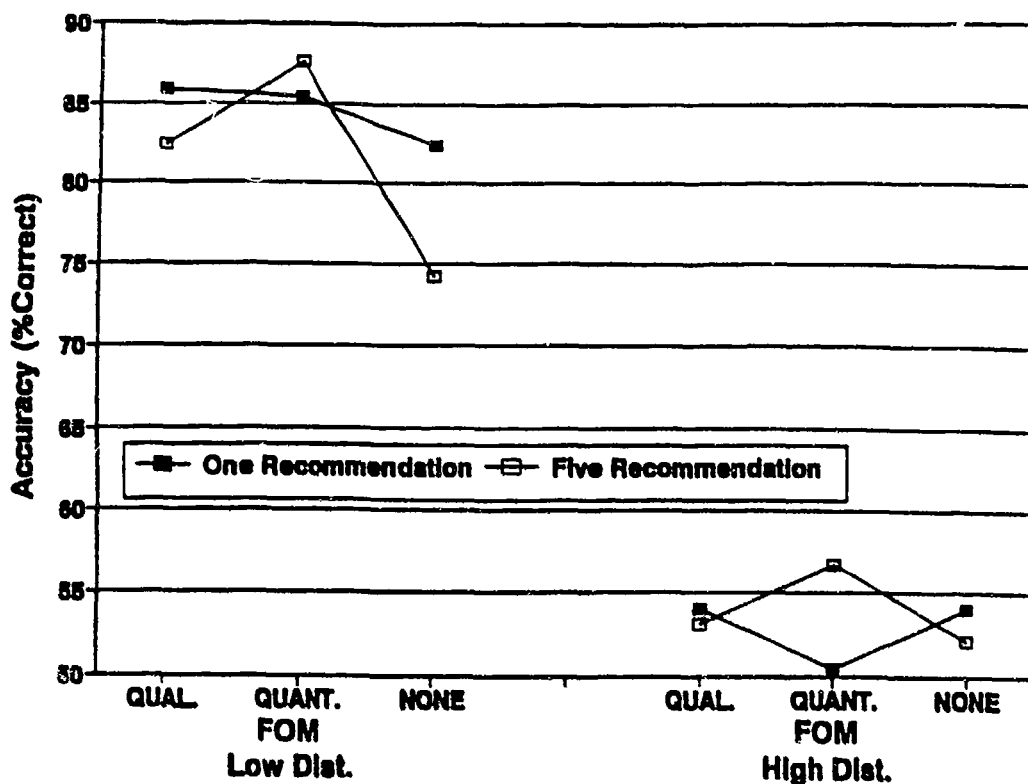


FIGURE 5. Accuracy Versus Display Type.

Errors by Type

The discussion above suggests that both the form of the FOM and the number of recommendations displayed influence the overall accuracy. However, a more detailed analysis of the subjects' errors provides additional insight into the decision processes involved.

Errors may be generally divided into two categories, type A errors that are the result of the subject's changing a correct ATR recommendation into an incorrect identification, and type B errors that are the consequence of the subject's accepting an incorrect ATR recommendation.

ATR Incorrect

Figure 6 shows the frequency with which subjects corrected incorrect ATR recommendations. This figure shows the results for study 1 and study 2, and the average for the two studies. An analysis of variance with three factors (number displayed, FOM, and distortion level) indicated that only the ND main effect for study 1 ($\alpha = 0.0015$) and the ND by DL interaction for study 2 ($\alpha = 0.0512$) were significant

for this performance measure. By individual contrasts, the ND effect at the high distortion level was significant for study 2 ($\alpha = 0.0286$) and marginally significant for study 1 ($\alpha = 0.0656$).

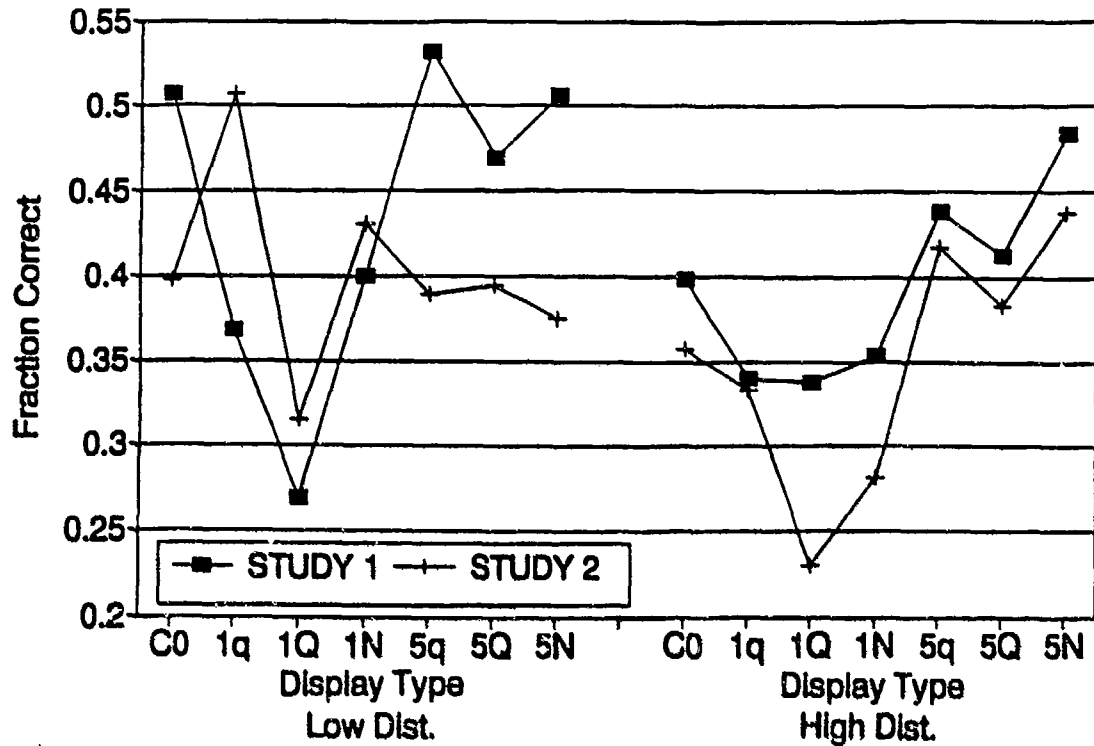


FIGURE 6. Fraction Correct Given ATR Wrong.

Increasing the number of recommendations from one to five increased the overall likelihood that an incorrect ATR recommendation would be corrected from 34.4 to 47.3% for study 1 and from 34.5 to 39.9% for study 2, for an overall average increase in accuracy of 9.1%. This increase was primarily due to the increase at the high distortion level. For the low distortion level, there was an increase of 15.6% for study 1 but a decrease of 2.8% for study 2, for an across study average increase of 6.4%. For the high distortion level the base values were 34.3 and 28.1%, respectively, with increases of 10.2 and 13.1% for an across study average increase of 11.6%.

It may also be observed from Figure 6 that the 1Q display, one recommendation and a quantitative FOM, gave the lowest likelihood of an incorrect recommendation being corrected for both distortion levels and both studies. When averaged across both studies and distortion levels, this difference between the 1Q type display and the remaining five displays was statistically significant ($\alpha = 0.0017$).

In summary, providing multiple recommendations increases the likelihood of an incorrect recommendation being corrected. An ATR that provides a single recommendation with a quantitative FOM minimizes the likelihood of incorrect recommendations being corrected.

ATR Correct

Figure 7 shows the frequency with which subjects changed correct ATR recommendations. For both studies, the type of FOM ($\alpha = 0.0339$ and $\alpha = 0.0022$), the number of recommendations displayed ($\alpha = 0.0010$ and $\alpha = 0.0002$), and the type of FOM by number of recommendations interaction ($\alpha = 0.0129$ and $\alpha = 0.0116$) significantly influenced this performance measure. For this measure, the DL by FOM interaction was significant in study 1 ($\alpha = 0.0176$) but not in study 2. At the low distortion level, the aided subjects altered 10.00% of the ATR's correct recommendations. This rate climbed to 23.92% for the high distortion level.

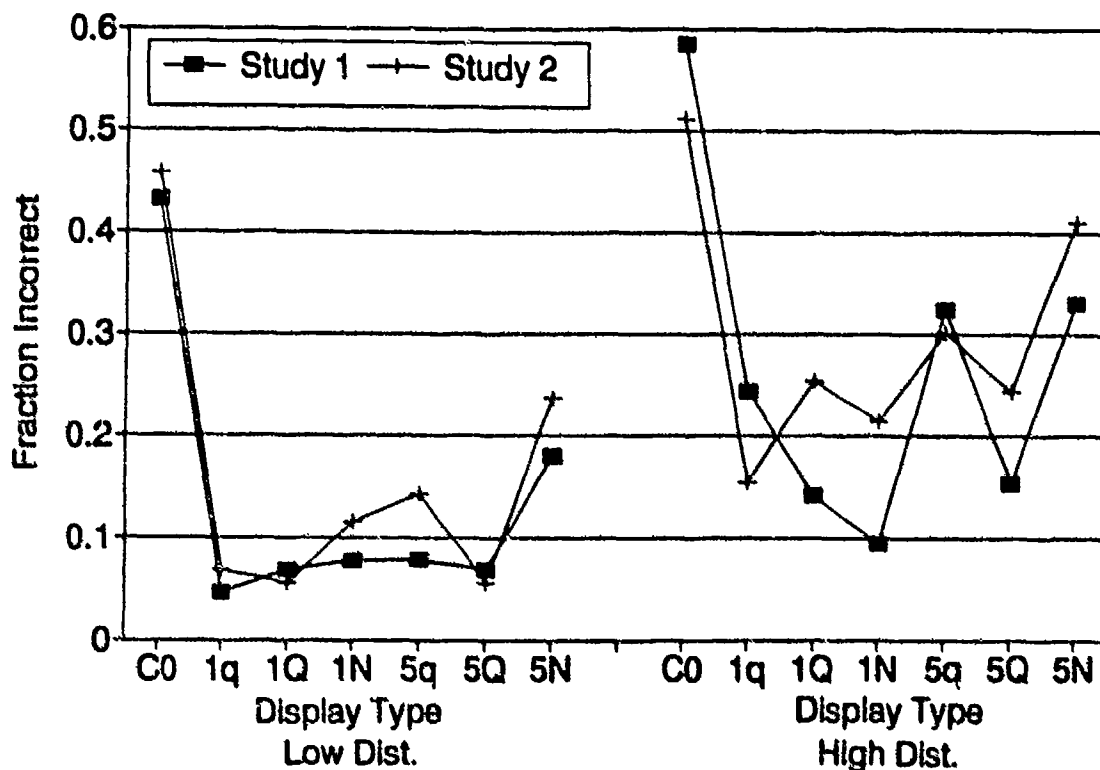


FIGURE 7. Fraction Incorrect Given ATR Correct.

When averaged across both distortion levels and studies, the three single recommendation display formats had nearly equal error rates, ranging from 12.6 to 13.0%. The five recommendation quantitative FOM display had a similar error rate of 13.1%. However, the remaining two formats had considerably higher rates, 21.2% for the five recommendation qualitative and 28.9% for the five recommendation no FOM. It was the significantly higher rates for these latter two display formats that caused both the main effects and interactions to be significant.

Imputed Reliabilities

Using signal detection theory (SDT), imputed ATR reliabilities were calculated for each display format and distortion level. The assumptions and details of these calculations are contained in the Appendix to this report. Imputed reliabilities are the ATR accuracies that are implied by the subjects' behavior. That is, the subjects were behaving as if the ATR's accuracy were the imputed value.

These imputed reliabilities are shown in Figure 8. Recall that at the low distortion level the ATR was correct in 84.94% of the trials. However when averaged across both studies, the imputed reliability was only 64.0%. Thus at the low distortion level, subjects were under-weighting the ATR's recommendations. In fact the 1Q display type had the highest implied reliability at this distortion level, which was only 78.8% in study 2.

Although not uniformly true, subjects tended to over-weight the ATR's recommendation at the high distortion level, where the imputed reliability averaged 46.7% and the actual was only 40%. While the three single recommendation formats averaged 55.3%, the 5q and 5N display formats under-weighted the ATR recommendation, averaging only 33.8 and 32.4%, respectively. The 5Q display format slightly over-weighted the recommendation, with an imputed reliability of 48.4%.

It should be noted that the imputed ATR reliabilities given above are conservative. For example, if one considered the likelihood of a recognized incorrect ATR recommendation being turned into a correct recommendation (which was 55.55% at the low distortion level), then the imputed reliabilities at this distortion level would drop to 29.6 and 28.9%, respectively, for the two studies.

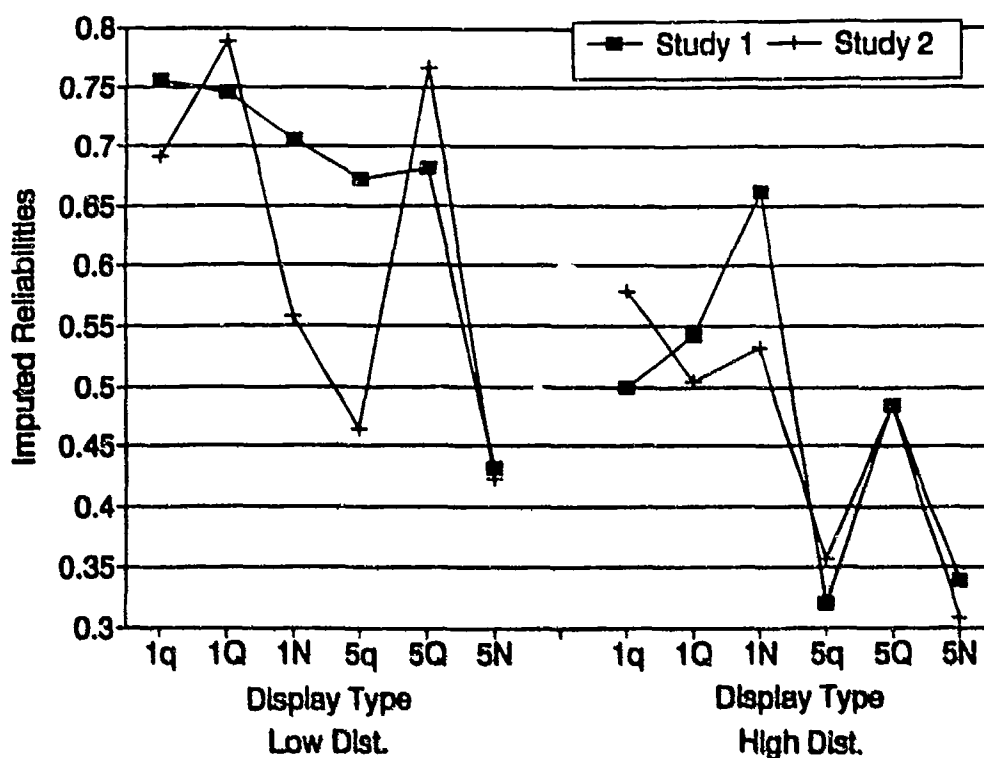


FIGURE 8. Imputed ATR Reliabilities.

Response Times

The number of recommendations displayed was the only factor that significantly influenced mean response times for both studies ($\alpha = 0.0014$ and $\alpha = 0.0010$). For study 2 both the FOM ($\alpha = 0.0387$) and the FOM by ND interaction ($\alpha = 0.0132$) were significant. Note that there were two design differences between the two studies. The bar graph was not functioning properly in study 1 and there was a 30-second time limit in study 2. Either of these two factors could possibly explain these observed differences.

Increasing the number of recommendations from one to five increased the mean response time, in both studies, as seen in Figure 9. For study 1 the mean response time increased from 15.5 to 18.7 seconds, an increase of 2.2 seconds. In study 2, the increase was smaller, again possibly due to the imposed time limitation. In this study, the mean response time increased from 10.9 to 12.0 seconds, a 1.1-second increase.

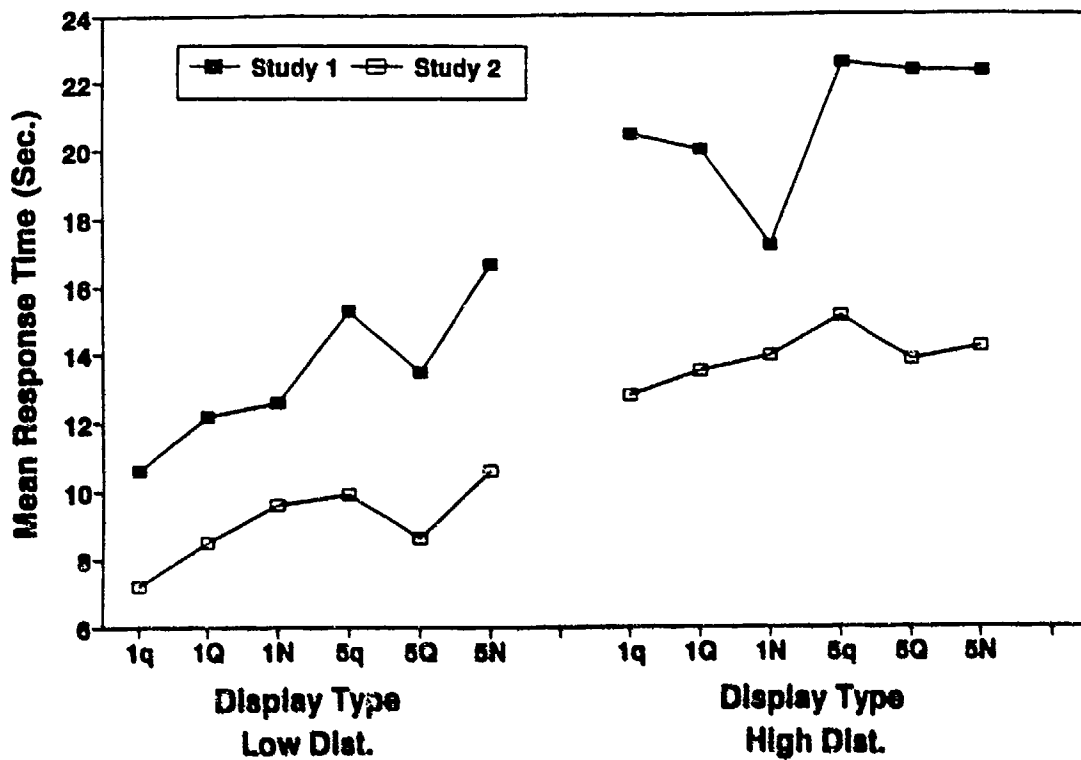


FIGURE 9. Mean Response Times Versus Display Type.

In study 2, the qualitative FOM resulted in both the fastest mean response time (10.0 seconds when paired with one recommendation) and the slowest (12.5 seconds when paired with five recommendations). The quantitative FOM had the smallest difference in mean response times between one and five recommendations in both studies. For the quantitative FOM, increasing the number of recommendations from one to five increased the mean response time from 11.0 to 11.2 seconds.

When the ATR provided five recommendations, the quantitative FOM resulted in the lowest mean response times for both studies, with mean responses times of 17.9 and 11.2 seconds, respectively. These mean response times were 1.0 and 1.2 seconds faster than the second fastest alternative.

SUMMARY AND DISCUSSION

In general, operator aiding by an ATR either reduces mean response time or increases accuracy. When an ATR is substantially more accurate than an operator, the system accuracy will approximate the ATR's accuracy, and the operator's response times will be substantially reduced. When the ATR and operator are approximately equally matched, the two will operate synergistically, without a substantive increase in response time. What will occur when the operator is substantially better than the ATR cannot be ascertained from this study.

An analysis of the operator's and ATR's errors indicated that they were using different identification algorithms that appeared to complement each other. If they had used similar algorithms, they would not have complemented each other and the system performance would not have exceeded the better of the two. Consequently, if there is to be an operator in an ATR system, the ATR algorithm designers may want to design an operator complementary algorithm.

Although, the ATR display format factors investigated in this study influenced both the likelihood that a correct ATR recommendation will be changed to an incorrect response and the likelihood that an incorrect recommendation will be converted to a correct recommendation, they had a much greater influence on the former than on the latter. In both studies, the number of recommendations provided, the type of FOM, and their interaction significantly influenced the likelihood that a correct ATR recommendation would be modified.

Increasing the number of recommendations displayed increased the likelihood that operators would disagree with the ATR, regardless of whether the ATR was correct.

If only one recommendation was provided, then the type of FOM included in the display had only a modest influence. However if the display format included multiple recommendations, the type of FOM became significant. With multiple recommendations, subjects were more likely to agree with an ATR's recommendation if the display contained a quantitative FOM, as opposed to a qualitative FOM or no FOM. The influence of the FOM type appears limited to the quantitative FOM paired with five recommendations. Under the conditions considered in this study, there appears to be little difference between a qualitative FOM and no FOM at all.

The effect of each of the various display formats investigated in the two studies can perhaps be best understood by examining the imputed ATR reliabilities. All formats resulted in the subjects' underestimating the ATR's reliability for the low distortion condition. The 5N condition (five recommendations with no FOM) was by far the worst. While the ATR's actual reliability was slightly over 85%, the imputed reliabilities for this condition were 43.07 and 42.24%, respectively, for the two studies. These values are approximately half the actual ATR's reliability. Even at the high distortion level, the 5N format resulted in imputed reliabilities of 34.06 and 30.91%, still substantially less than the actual figure of 40%.

RECOMMENDATION

This study has shown that there can be a significant interaction between an ATR and operators. ATR algorithms that have similar performance when operating alone can have significantly different system performance levels when coupled with an operator. Therefore, it is recommended that system operations be thoroughly tested with appropriately trained operators under either actual operating conditions or high fidelity simulations of those conditions.

Display formats that provide multiple recommendations without FOMs or with qualitative FOMs should not be used because they are associated with low imputed reliabilities. Quantitative FOMs should always be provided with multiple recommendations.

Because subjects tend to concur with ATR recommendations, either correct or incorrect, caution should be used in applications involving single recommendation quantitative FOM display formats. If the ATR is always going to be used under conditions where it will be highly reliable, then this display format would be appropriate. However in applications where the ATR's reliability can be expected to be moderate to low, this format would not be suitable.

ATR designers should consider the use of adaptive display formats. When the ATR is used under conditions where its first recommendation is expected to be highly reliable, only a single recommendation should be provided. Multiple recommendations should be provided when the ATR's reliability falls below a prescribed value.

Appendix
AGREEMENT-DISAGREEMENT AND
SIGNAL DETECTION THEORY

AGREEMENT-DISAGREEMENT

Consider the task facing an operator working with an ATR. First, the operator must, using the information available, decide whether the ATR's first recommendation is correct. If the subject concludes that the ATR's first recommendation is correct, then that recommendation is selected and the task is complete. If the subject concludes that the ATR's first recommendation is incorrect, then the subject must select the correct response, either from among the remaining recommendations or from the other choices. Thus the subject's task can conceptually be divided into two components: (1) deciding whether the ATR's first recommendation is correct, and (2) selecting the correct response if the answer to the first question is negative.

Consider the cases where the subjects disagreed with an incorrect ATR recommendation. The frequencies with which they then correctly identified the target are shown in Figure A-1. The distortion level was the most significant factor affecting this frequency, with $\alpha = 0.0329$ for study 1 and $\alpha = 0.0306$ for study 2. When low distortion level recommendations were averaged across both studies, 55.55% of the recognized incorrect ATR recommendations were corrected. Similarly, for the high distortion level, 45.7% were corrected. As a basis for comparison, the unaided classification accuracies were 54.0 and 39.7% for the low and high distortion levels, respectively.

The distortion level by number of recommendations interaction was significantly affected by this probability ($\alpha = 0.0039$) in study 2, but not in study 1. In study 1, the number of recommendations main effect was marginally significant ($\alpha = 0.0846$). An examination of the accuracies at each distortion level indicated that at the high distortion level, the addition of four recommendations increased the accuracy from 41.5 to 49.8%. At the low distortion level as the number of recommendations increased, the likelihood of a recognized incorrect recommendation being corrected increased in study 1, but decreased in study 2.

The likelihood of a subject agreeing with an erroneous ATR recommendation is shown in Figure A-2. This factor was significantly affected by the number of recommendations displayed, with $\alpha = 0.0001$ in study 1 and $\alpha = 0.0002$ in study 2. Both the type of FOM and the FOM by ND interaction were significant in study 2 ($\alpha = 0.0022$ and $\alpha = 0.0116$), although neither was significant in study 1.

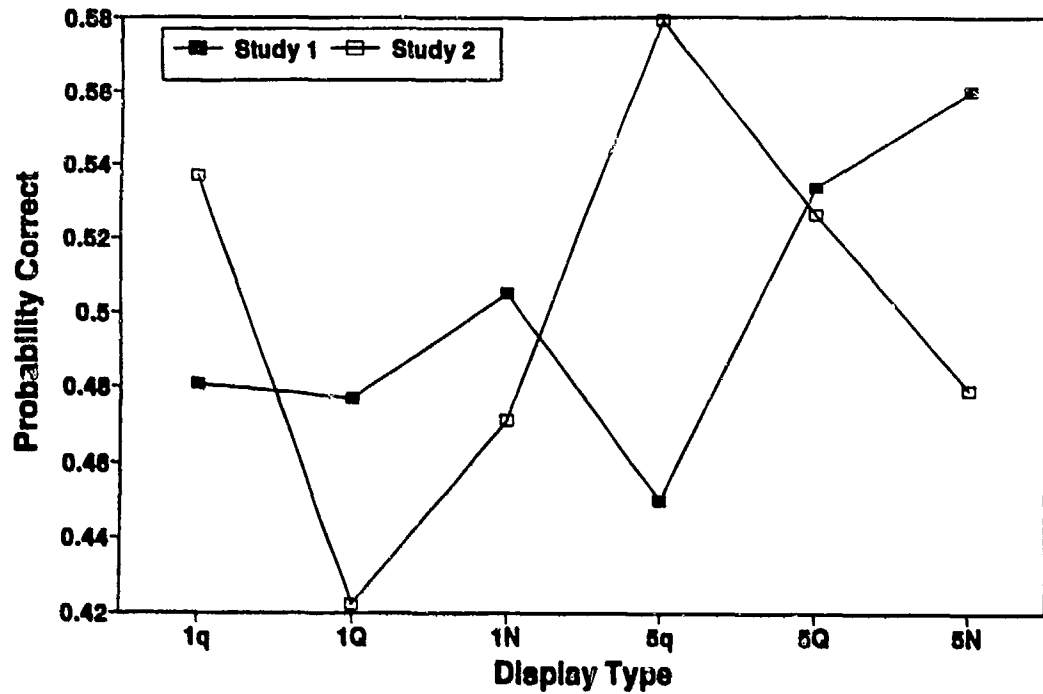


FIGURE A-1. Probability Subject Correct When ATR Wrong and Subject Disagrees.

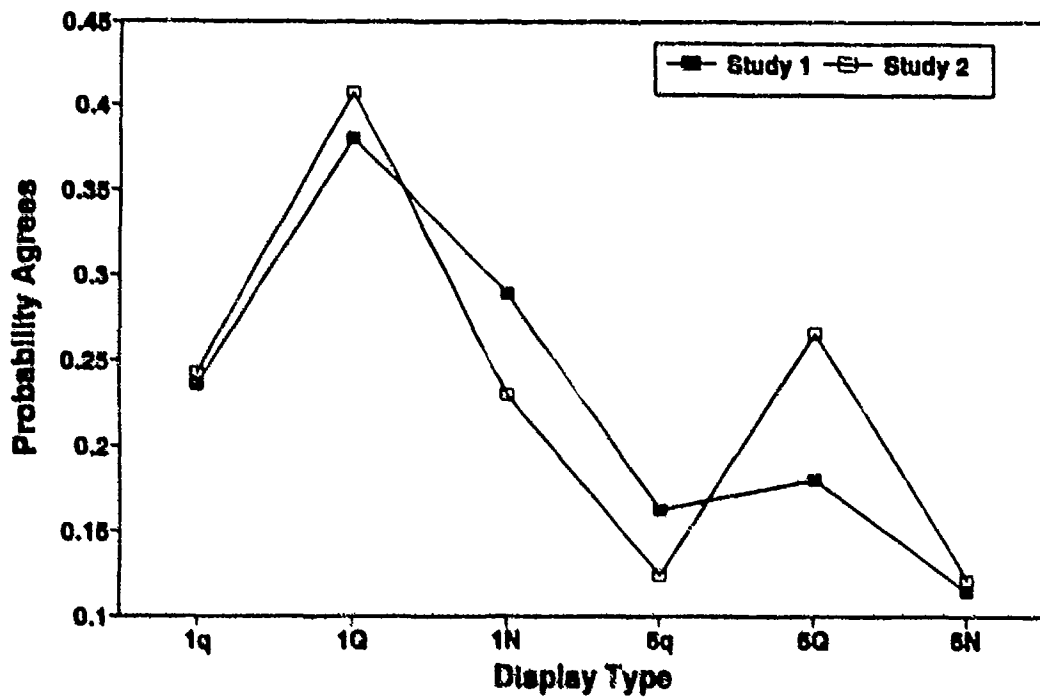


FIGURE A-2. Probability Subject Agrees When ATR Wrong.

The distortion level factor also significantly affected this probability in both studies with $\alpha = 0.0061$ in study 1 and $\alpha < 0.0001$ in study 2. The likelihood that a subject would agree with an erroneous ATR recommendation decreased from 26.9% at the low distortion level to 18.9% at the high distortion level.

Signal Detection Theory

The general theory of signal detectability (SDT) was originally developed by Peterson, Birdsall, and Fox (Reference 8) and by Van Meter and Middleton (Reference 9). They developed a general theory for the detectability of a signal superimposed over a noise background. This general theory of signal detectability was then adapted to psychophysics by Tanner and Sweets (Reference 10) and others.

In the Tanner and Sweets paradigm, the likelihood that a signal will be detected is characterized by the signal to noise ratio, d' . For example, a signal to noise ratio of 2.00 indicates a signal strength equal to twice the standard deviation of the background noise.

If the signal to noise ratio is not large, then errors are inevitable and the subjects must decide how to balance the error of reporting a signal when none is present and the error of reporting no signal when one is present. This error trade-off is usually described by β , the ratio of the likelihood of a signal plus noise to the likelihood of only noise at the cutoff or critical point. One of the attractive features of SDT is that it effectively separates the error trade-off from the signal strength. The interested reader is referred to Reference 11 for a more detailed discussion of SDT.

The SDT paradigm can be adapted to the current identification task if a correct ATR recommendation is thought of as a signal with noise and an incorrect recommendation is thought of as noise alone. Thus, the subject's decision to agree or disagree with the ATR can be considered the analog of the decision to report or not report a signal.

Following this approach allows the calculation of a d' and a β for each ATR display format for each of the studies. These are shown in Figures A-3 and A-4, respectively. For the low distortion level, the average signal to noise ratio was 2.0323 for study 1 and 1.9661 for study 2, for an overall average of 1.9992. The high distortion level gave average d' s of 1.8045 and 1.509, respectively, for the two studies, for an overall average of 1.6568.

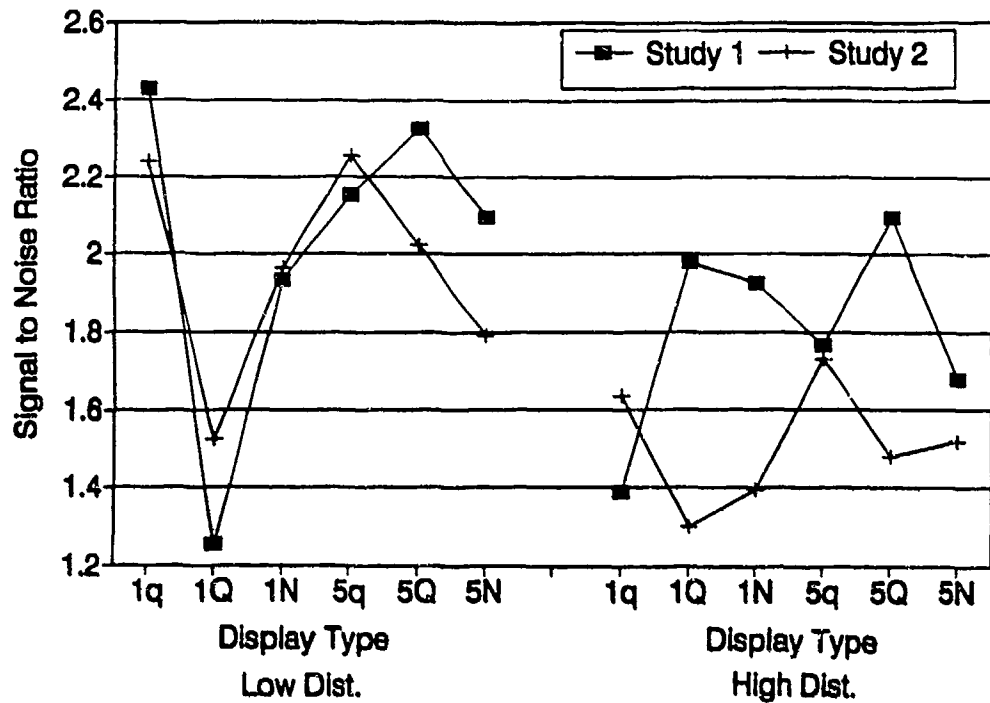


FIGURE A-3. d' Signal to Noise Ratio.

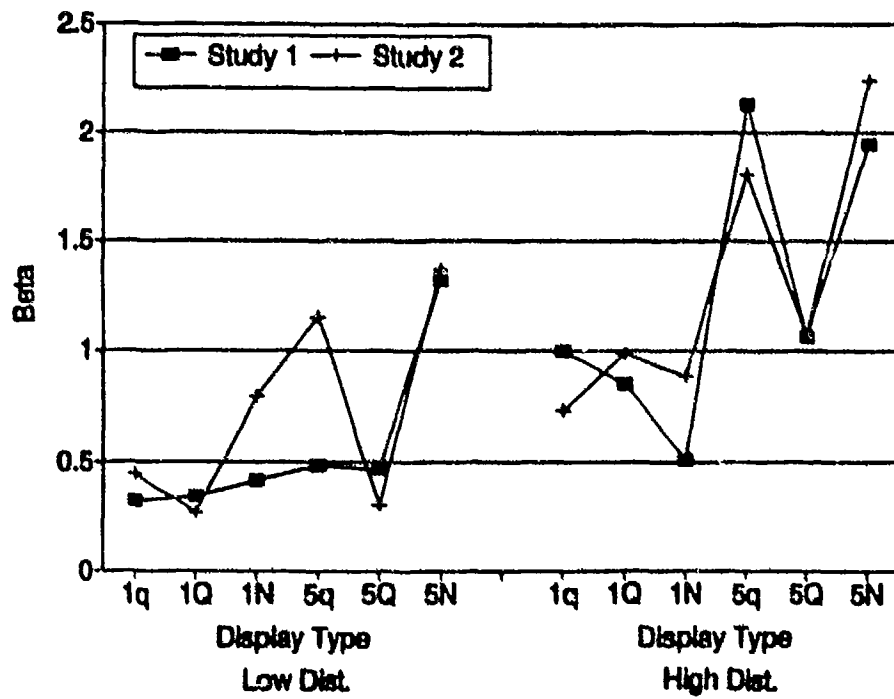


FIGURE A-4. Beta.

With the exception of the 1Q display type at the low distortion level, there was little variation in the d' values. The d' values ranged from 1.390 to 2.330 for the low and from 1.510 to 1.789 for the high distortion level. The 1Q display type signal to noise ratio, which was 1.390 at the low and 1.640 at the high distortion levels, was the lowest of the six display types in three out of the four possible cases.

The β values averaged across both studies were 0.6404 and 1.2661 for the low and high distortion levels, respectively. An inspection of Figure A-4 reveals several anomalies. First, the five recommendation displays' β s were consistently greater than the one recommendation displays. Second, with the exception of the high distortion level for study 1, the 5N display format's (five recommendation no FOM) β was considerably higher than the others, averaging 1.7171. The 5q display format was the second highest, with an average of 1.3931. The remaining four formats (1q, 1Q, 1N, and 5Q) averaged 0.6246, 0.6094, 0.6500, and 0.7254, respectively.

If the relative costs of the various combinations of reports (agree, disagree) and conditions (ATR correct, ATR wrong) are known, an optimal β can be calculated. Let β^* represent the optimal value of β . Then

$$\beta^* = \frac{V_{dw} + C_{aw}}{V_{ac} + C_{dc}} * \frac{P(ATRW)}{P(ATRC)}$$

where

V_{dw} = the value of disagreeing with an erroneous ATR recommendation

V_{ac} = the value of agreeing with a correct ATR recommendation

C_{aw} = the cost associated with agreeing with an erroneous recommendation

C_{dc} = the cost associated with disagreeing with a correct ATR recommendation

$P(ATRW)$ = the probability of an incorrect ATR recommendation

$P(ATRC)$ = the probability the ATR recommendation is correct.

If the costs are assumed to be equal, then the equation above reduces to

$$\beta^* = \frac{P(ATRW)}{P(ATRC)}$$

or letting $P(ATRC) = r$,

$$\beta^* = (1-r)/r$$

Inverting this equation gives

$$r = 1/(1+\beta^*)$$

This equation can be used to calculate imputed probabilities (reliabilities). These imputed reliabilities are shown in Figure 8 for each distortion level and ATR display format.

REFERENCES

1. Naval Weapons Center. *Targeting Decisions Using Multiple Imaging Sensors: The Use of Image Quality Figures of Merit*, by Scott A. Weisgerber, Marion P. Kibbe, and Carrie F. Quesnell. China Lake, Calif., NWC, July 1991. (NWC TP 7123, publication UNCLASSIFIED.)
2. ----- . *Targeting Decisions Using Multiple Imaging Sensors: Operator Performance and Calibration*, by Marion P. Kibbe and Scott A. Weisgerber. China Lake, Calif., NWC, February 1990. (NWC TP 7054, publication UNCLASSIFIED.)
3. ----- . *Multisensor Evaluation Framework*, by David C. Foyle. China Lake, Calif., NWC, September 1989. (NWC TP 7027, publication UNCLASSIFIED.)
4. ----- . *Presentation Format as a Determinant of Targeting Task Performance*, by Steven R. Adams. China Lake, Calif., NWC, May 1991. (NWC TP 7125, publication UNCLASSIFIED.)
5. ----- . *Operator Use of Automated Target Recognition (ATR) Systems Under Time Constrained Conditions*, by Marion P. Kibbe and Scott A. Weisgerber. China Lake, Calif., NWC, December 1991. (NWC TP 7136, publication UNCLASSIFIED.)
6. ----- . *Operator Ship Classification Using an Automatic Target Recognition (ATR) System in Conjunction With Forward-Looking Infrared (FLIR) Imagery*, by Scott A. Weisgerber and Susan F. Savage. China Lake, Calif., NWC, November 1990. (NWC TP 7101, publication UNCLASSIFIED.)
7. *Jane's Fighting Ships*, Capt. John Moore, ed. Royal Navy, London, England, Jane's Publishing Co., 1982.
8. Peterson, W. W., Birdsall, T. G., and Fox, W. C. "The Theory of Signal Detectability," in *Inst. Radio Engrs. Trans. Professional Group on Information Theory*, 1954, PGIT-4, pp. 171-212.
9. Van Meter, D., and Middleton, D. "Modern Statistical Approaches to Reception in Communication Theory," in *Inst. Radio Engrs. Trans. Professional Group on Information Theory*, 1954, PGIT-4, pp. 119-145.

10. Tanner, W. P., and Sweets, J. A. "A Decision-making Theory of Visual Detection," in *Psychological Review*, Vol. 61, pp. 401-409.
11. Welford, A. T. *Fundamentals of Skill*. Methuen and Company, LTD, London, 1968.

INITIAL DISTRIBUTION

- 4 Naval Air Systems Command
AIR-5313 (2)
AIR-546TG (1)
AIR-5462 (1)
- 3 Chief of Naval Research, Arlington
OCNR-10 (1)
OCNR-20 (2)
- 1 Air Test and Evaluation Squadron 5, China Lake (Technical Library)
- 1 Naval Air Warfare Center, Aircraft Division, Indianapolis (Technical Library)
- 2 Naval Air Warfare Center, Aircraft Division, Patuxent River
Human Factors Branch (1)
Technical Library (1)
- 3 Naval Air Warfare Center, Aircraft Division, Warminster
Code 6021 (1)
Code 6022 (1)
Technical Library (1)
- 1 Naval Air Warfare Center Weapons Division, Point Mugu (Technical Library)
- 3 Naval Command Control and Ocean Surveillance Center RDTE Division, San Diego
Grossman (1)
Human Factors Group (1)
Technical Library (1)
- 1 Naval Health Research Center, San Diego (Technical Library)
- 4 Naval Postgraduate School, Monterey
J. Lind (3)
Technical Library (1)
- 1 Naval Research Laboratory (Technical Library)
- 1 Naval Surface Warfare Center, Dahlgren Division, Dahlgren (Technical Library)
- 1 Naval Surface Warfare Center, Dahlgren Division Detachment White Oak, Silver Spring (Technical Library)
- 1 Naval War College, Newport (Technical Library)
- 3 Office of Naval Technology, Arlington (ONT-222, Dr. S. Collyer)
- 1 Army Training and Doctrine Command, Fort Monroe (Technical Library)
- 1 Headquarters, U. S. Army, Fort Belvoir (Technical Library)
- 1 Army Aeromedical Research Laboratory, Fort Rucker (Technical Library)
- 1 Army Human Engineering Laboratory, Aberdeen Proving Ground (Technical Library)
- 1 Army Material Systems Analysis Activity, Aberdeen Proving Ground (Technical Library)
- 1 Headquarters, Air Force Materials Command, Wright-Patterson Air Force Base (Technical Library)
- 1 Tactical Air Command, Langley Air Force Base (TAC/DRIY, Technical Library)
- 2 Wright Laboratory, Dynamics Directorate, Wright-Patterson Air Force Base
WL/AAND-1 (1)
Technical Library (1)
- 1 Defense Intelligence Agency (Technical Library)
- 2 Defense Technical Information Center, Alexandria
- 1 Institute for Defense Analyses, Alexandria, VA (Technical Library)

OW CENTER DISTRIBUTION

1 Code C02	1 Code C2107
1 Code C02B	1 Code C211
1 Code C02B01	1 Code C215
1 Code C023	1 Code C2151
1 Code C024	10 Code C2152, Kibbe
1 Code C0242	1 Code C2153
1 Code C024205	1 Code C2156
1 Code C02421	1 Code C2158
1 Code C03A	1 Code C2159
1 Code C21	1 Code C219
1 Code C2101	1 Code C2587, GIDEP
1 Code C2103	1 Code C27
1 Code C2104	4 Code C643 (3 plus Archives Copy)